EVALUATING MUSIC SEQUENCE MODELS THROUGH MISSING DATA

Thierry Bertin-Mahieux[*], *Graham Grindlay*

Columbia University
LabROSA
New York, USA

Ron J. Weiss[†] *and Daniel P.W. Ellis*

New York University / Columbia University
MARL / LabROSA
New York, USA

ABSTRACT

Building models of the structure in musical signals raises the question of how to evaluate and compare different modeling approaches. One possibility is to use the model to impute deliberately-removed patches of missing data, then to compare the model's predictions with the part that was removed. We analyze a corpus of popular music audio represented as beat-synchronous chroma features, and compare imputation based on simple linear prediction to more complex models including nearest neighbor selection and shift-invariant probabilistic latent component analysis. Simple linear models perform best according to Euclidean distance, despite producing stationary results which are not musically meaningful. We therefore investigate alternate evaluation measures and observe that an entropy difference metric correlates better with our expectations for musically consistent reconstructions. Under this measure, the best-performing imputation algorithm reconstructs masked sections by choosing the nearest neighbor to the surrounding observations within the song. This result is consistent with the large amount of repetition found in pop music.

Index Terms— Missing data imputation, music audio, chroma features, entropy difference, music sequence models

1. INTRODUCTION

As with many classes of time-series data, musical signals contain substantial amounts of complex structural information. Given the intricate nature of this structure, unsupervised modeling of music is a particularly alluring yet challenging task. Precise models based on large music archives could have a great impact, not only on musicology but also in numerous commercial applications, including recommendation systems, digital rights management, and creative tools. We are particularly interested in models that capture local patterns (or "patches") in the data. A successful method for describing music as a collection of patches would be useful for tasks such as song similarity (recognizing songs with similar patterns), song segmentation (labeling chorus/verse structure), and cover song recognition (identifying songs with similar high-level patterns). All of these tasks would benefit from patch models that capture high-level musical characteristics while remaining faithful to the observed signal data. It is unclear, however, how to evaluate the extent to which a model captures "musically sensible" structure. This is the question addressed in the current work through the task of missing data imputation, and via a number of metrics which are sensitive to musically meaningful aspects of the data.

Imputation refers to the problem of filling in data items that have been lost or deleted. Previous work on audio-related applications has included speech denoising [1, 2], source separation [3], bandwidth expansion [4], and model evaluation [5]. Much of this work has focused on small deletions such as a single time-frame. The relatively slowly-varying properties of music allow single-frame deletions to be imputed with good accuracy based on simple extrapolation. However as the amount of missing data increases, the problem becomes significantly more challenging. We therefore focus on the task of *multi-frame* imputation in which long segments of the signal are missing. Obtaining good performance on this task requires the use of models that effectively exploit longer-term musical structure. The task therefore serves as a good way to more carefully evaluate the temporal aspects of competing models.

The difficulties of multi-frame imputation of music data are illustrated in Figure 1. For instance, linear prediction, which we found to give the smallest Euclidean distances over a large test set, is used to impute the missing data as shown in the 6th row.Visually, it is evident that linear prediction yields an overly-smooth reconstruction and is unable to restore any temporal evolution within the missing section. A more musically-coherent representation demands a model better that is able to employ long-span information, such as knowledge of repetitions within a song. Unfortunately, simple measures like Euclidean distance are not sensitive to musical coherence, and will lead to solutions like linear prediction. In this paper, we argue for a more extensive set of metrics, able to properly evaluate a model's ability to predict musical sequences.

2. TASK DEFINITION

2.1. DATA AND FEATURES

We use a set of 5000 songs taken from the *morecowbell.dj* dataset [6], which consists of a wide variety of (predominantly western) pop music. The raw audio is converted to a chromagram representation using the online Echo Nest API.[1] A chromagram is similar in spirit to a constant-Q spectrogram except that pitch content is folded into a single octave of 12 discrete bins, each corresponding to a particular semitone (e.g. one key on a piano).

Music naturally evolves over a time scale expressed in beats, so instead of the fixed-length frames commonly used in other audio processing, we form beat-aligned chroma by resampling a chromagram

[*]supported in part by a NSERC PG scholarship.

[†]supported by NSF grant IIS-0844654 and Institute of Museum and Library Services grant LG-06-08-0073-08.

This work is supported by NSF grant IIS-0713334 and by a gift from Google, Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

[1] `http://developer.echonest.com`

to make each chroma (column) vector span a single beat. This representation has been successfully used for cover song recognition [7], segmentation [8], and in our previous work on patch modeling [6]. Loudness variation is removed by normalizing each column to have a maximum value of one.

2.2. ALGORITHMS

To better understand the multi-beat imputation task, we present several benchmark algorithms of varying sophistication. While the simpler algorithms cannot be expected to provide high-quality imputations, their relative performance is of interest. Simple methods include filling the missing chroma bins with draws from a uniform $[0, 1)$ distribution (Figure 1, 3rd row), or using a beat picked at random from elsewhere in the song, or using the average of all remaining beats.

Our first trained model uses a linear transform to predict the next beat from the previous N beats. As it explicitly learns to minimize Euclidean distance on the remaining data, this method performs very well under certain metrics (see Figure 1). From our experiments, $N = 1$ or $N = 2$ works best.

Nearest neighbor (1NN) is a promising technique given the repetitious structure of songs. By looking at the beats near to the missing data, we can impute a reconstruction by spotting and copying a similar neighborhood. Note that instead of using the remaining part of the song, one could also use a codebook constructed from other songs, which could involve kNN with $k > 1$.

The algorithms mentioned so far rely on low-level information which is unlikely be sufficient for a fully satisfactory multi-frame imputation. Therefore, we also try shift-invariant probabilistic latent component analysis (SIPLCA) [4, 8], which is able to capture more subtle temporal dependencies. SIPLCA extracts template chroma patches that are consistently repeated throughout a song. Missing data is imputed by identifying templates consistent with the features neighboring the missing segment. In the case of missing segments longer than the identified templates, our imputation algorithm utilizes a hierarchical reconstruction scheme, whose detailed explanation is beyond the scope of this paper. SIPLCA for imputation is ongoing research and we refer the interested reader to our code for the moment.

2.3. EVALUATION METRICS

Euclidean distance is a natural first choice for evaluating reconstruction and encoding tasks. However, as can be seen in Figure 1, algorithms which minimize this metric do not necessarily capture all of the relevant data statistics. In particular, the solution is overly smooth and longer-term temporal patterns are poorly reconstructed.

Clearly, the simplicity of the model is largely responsible for these issues. However, inherent in the Euclidean distance criterion is a preference for smooth solutions. To see why, consider the toy example of Figure 2b. We approximate the original one-dimensional signal, a square wave (solid line), by two signals: a translated identical square wave (dot-dashed line), and a constant signal (dashed line). The first signal has an average reconstruction error of 0.50 using Euclidean distance. The second signal has an average error of only 0.25, despite the fact that it does not appear to reflect the overall shape of the original data.

The class of Minkowski distances are given by the form $d_p = |x_1 - x_2|^p$, of which the Euclidean is a special case ($p = 2$). In general, as $p \to 0$ the resulting distance measures penalize small values more heavily. This has the effect of favoring "sharper" data sequences and is illustrated in Figure 2a. The greyed rectangle represents the case where a reconstruction is considered valid (error equals 0 if it is between some δ of the original and wrong otherwise (error equals 1). The Minkowski distance approximates this case when both $\delta \to 0$ and $p \to 0$. In the experiments described below, we consider $p = 0.5$ and $p = 2$. With $p = 0.5$, on the example of Figure 2b, the translated square wave is now favored (with a reconstruction error of 0.5) to the second signal whose reconstruction error is now 0.71.

Looking at the differences between the original data and the imputed values, it appears that we need to encourage sparsity in our solutions. Entropy is a natural measure of the sparsity of a distribution and therefore it makes sense to consider related metrics. We examined the (symmetric) Kullback-Leibler divergence, (approximate) conditional entropy [9], Jensen difference [10], and the normalize difference entropy (D-ENT) [11]. The Jensen difference of two distribution x_1 and x_2 is given by

$$J(x_1, x_2) = H\left(\frac{x_1 + x_2}{2}\right) - \frac{H(x_1) + H(x_2)}{2}$$

where $H(x)$ denotes the entropy of x. For D-ENT, we build a 10-bin histogram of feature values, and compute the difference of entropy for each bin between the original distribution and its approximation:

$$\text{D-ENT}(\mathbf{b}_1, \mathbf{b}_2) = \frac{1}{10}\sum_{i=1}^{10} \frac{\log_2 b_{1,i} - \log_2 b_{2,i}}{\log_2 b_{1,i}}$$

where $\mathbf{b}_1$ is the set of bins for x_1. Note that D-ENT is not symmetric. In our experiments, we set the first vector to be the original features. Also, the position of the different values does not matter.

Another idea to measure the "sharpness" of a reconstruction compared to the original using the first-order difference of the feature vectors across time. We compute the delta difference by summing the absolute value of the delta, then taking the absolutec difference. Once again, the position of the delta values does not matter.

We explained above why Euclidean distance can justify disappointing reconstructions. At the same time, we do not argue that we should ignore or replace it. Euclidean distance measures reconstruction in a fundamental way. We believe we need a combination of measures to quantify the quality of music patterns, Euclidean distance being one of them. In the next section, we investigate which measures behave in a similar way, thus helping us to decide which ones are useful.

3. EXPERIMENTS

As we mentioned in Section 2.3, there exist numerous metrics to evaluate the reconstruction of a signal. Euclidean distance is a natural first choice, but it does not necessarily reward musical structure. In order to justify our set of reported measures, we computed the statistical similarity between a number of distance metrics. We use Pearson's correlation which is defined as:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Note that $-1 \leq \rho_{X,Y} \leq 1$, and an absolute value close to 1 means high correlation. We compute $\rho_{X,Y}$ for all pairs of error measures using the results obtained by the three imputation methods described in Subsection 2.2 on a two multi-frame imputation task. The results are shown in Table 1. Note that we experimented with a number of other distance metrics, including symmetric Kullback-Leibler divergence, conditional entropy, and cosine distance. Although we do not

	Euclidean	$d_{1/2}$	Jensen	Delta diff.	D-ENT
Euclidean	1				
$d_{1/2}$	0.90	1			
Jensen	0.84	0.71	1		
Delta diff.	**0.45**	0.52	0.38	1	
D-ENT	**0.12**	0.27	0.20	0.52	1

Table 1: Pearson correlation between measures computed using the full dataset with several imputation schemes and amounts of missing data. Correlation is symmetric.

method	Euclidean	delta diff.	D-ENT
random	0.168	0.135	0.252
average	0.079	0.180	0.430
1NN	0.072	**0.028**	**0.123**
lin. trans.	**0.056**	0.170	0.479
SIPLCA	0.060	0.149	0.395

Table 2: Results on 15 missing beats by different methods on 5000 songs and measured using Euclidean distance, delta difference, and D-ENT.

include detailed results due to lack of space, all of these metrics were highly correlated with Euclidean distance.

It is worth pointing out that the delta difference and D-ENT metrics differ from the rest in that they compare patches in a holistic manner. This stands in contrast to the other measures which work in a point-by-point fashion. Additionally, these two metrics measure a fundamentally different quality of the patch data in that they assess patch "granularity" as opposed to individual chroma bin levels.

We take a closer look at the divergence between two of the measures (Eucliean and D-ENT) in Figure 3. This shows the performance of three methods for different numbers of missing beats. As we can see from Table 1, they have a low empirical correlation of 0.12. The nearest neighbor (NN) method creates a reconstruction with a granularity similar to the original for all mask sizes. This implies that D-ENT is approximately constant throughout the songs and imputation using NN can easily preserve this characteristic. Tthe linear transform successfully learns to minimize the Euclidean distance. However, as is the case with D-ENT, the linear transform results in a substantial amount of smoothing (see the 6th row of Figure 1). The average reconstruction has the same D-ENT as the linear transform which is consistent with Figure 1. However, due to less smoothing (or perhaps less intelligent smoothing), it does not do as well in terms of Euclidean distance.

Figure 4 shows another example of imputation using different algorithms. The algorithms are ordered according to Euclidean distance. However, we can see that in this case, delta difference would have ranked NN first, followed by SIPLCA whose reconstructions seem more faithful to the original.

Table 2 shows the results of 15-beat imputation using 5000 songs. The linear transform is a clear winner based on Euclidean distance. As before, nearest neighbor's strength is to preserve the texture of the original patch as can be seen from the D-ENT score. We do not present all possible results (different numbers of missing beats, other error measures, etc.) due to space constraints, but given many of the other measures' high correlations to Euclidean distance, the differences are generally small.

4. CONCLUSION

We investigate the task of multi-frame imputation as a method for evaluating models of music sequences. Key to this evaluation is the definition of appropriate performance metrics. Experimental results over a data set of thousands of songs demonstrate that many standard metrics for comparing feature sequences, including Euclidean distance and Kullback-Leibler divergence, do not reliably measure the ability of an imputation algorithm to produce musically consistent reconstructions. We therefore propose to complement Euclidean distance with a measure of the entropy difference between the original features and their reconstruction. The proposed measure more consistently predicts an algorithm's ability to produce musically coherent reconstructions that are consistent with the original signal. The best performing imputation algorithm according to Euclidean distance often produces poor reconstructions, preferring to reuse a single sustained chord. The same linear prediction model performs poorly under the proposed measures, while more sophisticated sequence-based models show significantly better performance.

Given an appropriate framework for evaluating music sequence models, we intend to shift our focus to the exploration of more sophisticated sequential imputation algorithms in the future, including hidden Markov models and SIPLCA. Our goal is to encourage researchers to further explore this task. We have therefore made the code and data to reproduce these results available online.[2]

5. REFERENCES

[1] M. Cooke, A. Morris, and P. Green, "Recognising occluded speech," in *ESCA ETRW on The Auditory Basis of Speech Recognition*, Keele, 1996.

[2] B. Raj, R. Singh, and R.M Stern, "Inference of missing spectrographic features for robust speech recognition," in *International Conference on Spoken Language Processing*, 1998.

[3] M. Reyes-Gomez, N. Jojic, and D. Ellis, "Deformable spectrograms," in *Proc. Artificial Intelligence and Statistics*, 2005.

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *IEEE Workshop in Machine Learning for Signal Processing (MLSP)*, 2009.

[5] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. International Conference on Machine Learning (ICML)*, 2010, pp. 641–648.

[6] T. Bertin-Mahieux, R. Weiss, and D. Ellis, "Clustering beat-chroma patterns in a large music database," in *Proc. International Conference on Music Information Retrieval*, 2010.

[7] D. Ellis and G. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[8] R. Weiss and J. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, 2010.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1226–1238, 2005.

[2] http://www.columbia.edu/~tb2332/proj/imputation.html

[10] O. Michel, R.G. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measures," in *Proc. IEEE International Symposium on Time-Frequency and Time-Scale Analysis*, 1994, pp. 64–67.

[11] M. Mentzelopoulos and A. Psarrou, "Key-frame extraction algorithm using entropy difference," in *Proc. ACM International Workshop on Multimedia Information Retrieval*, 2004.

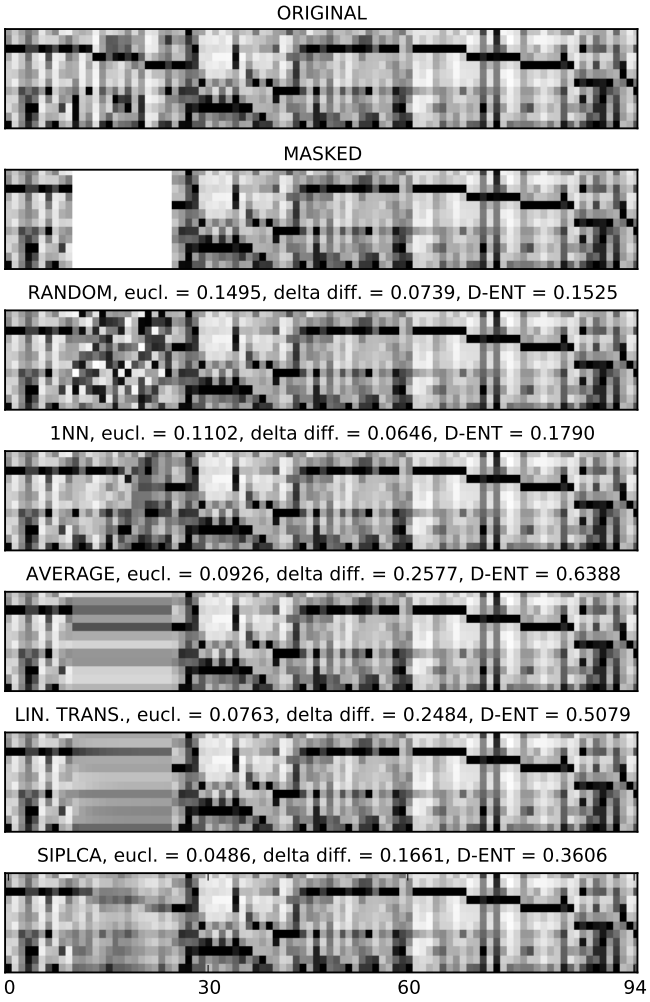

Fig. 1: 15-beat imputation example. Rows are (1) original data, (2) masked data, (3) random reconstruction, (4) nearest-neighbor reconstruction, (5) reconstruction using average of nearby beats, (6) reconstruction using linear transform of one previous beat, and (7) SIPLCA (see Section 2.2 for algorithmic details). Note that only excerpts are shown.

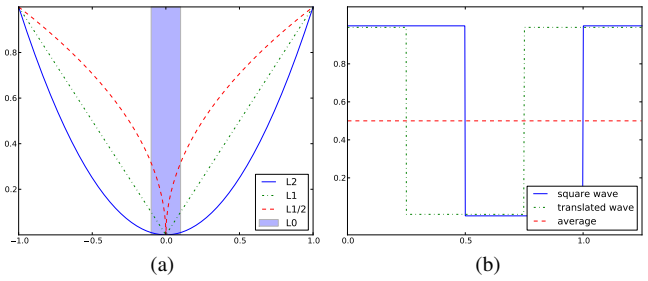

Fig. 2: (a) Effect of different measures on one-dimensional data. (b) Reconstruction error between a square wave and two approximations, a square wave translated by a quarter of the period, and the average function. Average error between original and translated wave is always 0.5 for any Minkowski measure d_p on $[0, 1]$. For the average function, the errors are 0.25, 0.5 and 0.71 for d_2, d_1 and $d_{1/2}$ respectively.

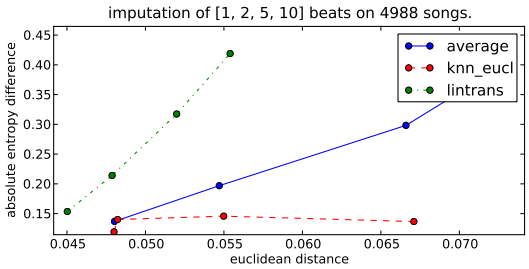

Fig. 3: Reconstruction error as a function of imputation algorithm and number of masked beats. The error metrics shown are D-ENT and Euclidean distance. In all cases, Euclidean distance increases with the number of masked beats. Lower left is better.

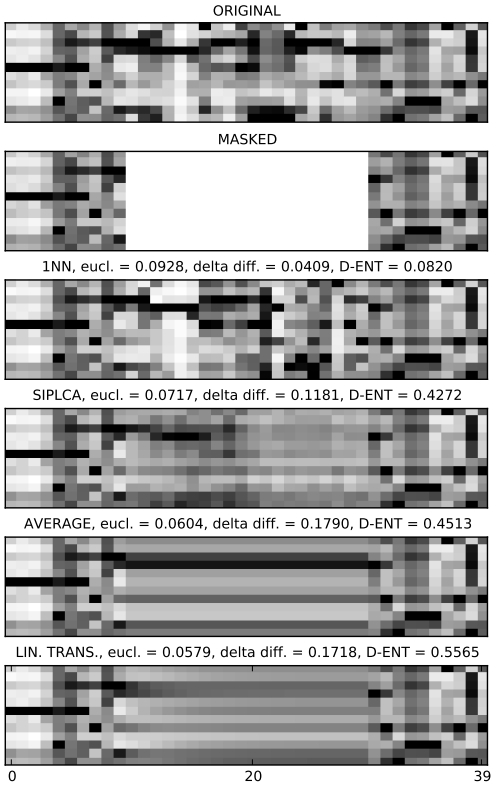

Fig. 4: 20 beats imputation example, rows are 1) original 2) original masked, then reconstruction using same algorithms as Fig 1 (except random).